Word Structure and Tone Perception in Mandarin

Hansjörg Mixdorff, Yu Hu***

*Faculty of Computer Science and Media, TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**Man Machine Voice Communication Laboratory, University of Science and Technology of China, Hefei, China
yuhu@iflytek.com

Abstract

This paper presents results concerning the relationship between word structure in terms of number of syllables and tonal realization in Mandarin. It examines whether the fact that a word (in our context a prosodic word) is more complex implies certain tonal reductions. Our hypothesis is that a monosyllabic word will be uttered more carefully than a polysyllabic word due to the potentially larger number of possibly confusable words. We also examine whether the total number of syllables in a word has an effect, creating more tonal reductions in longer than in shorter words. A database of Mandarin originally designed for concatenative speech synthesis and segmented into prosodic words was statistically analyzed regarding the occurrences of syllable/tone combinations in prosodic words of varying length. 10 sets of syllables were selected comprising all four tones of Mandarin and occurring as monosyllabic words as well as in varying positions in two- to five-syllable prosodic words. The target syllables were then extracted from their original context and presented to native speakers of Mandarin who had to decide which tone they perceived. The results of the perception test indicate, inter alia, that perception of syllables taken from polysyllables indeed is more error prone than that of monosyllabic words. The number of syllables in a word, however, has only a weak influence. Furthermore, reductions mostly appear for syllables in certain locations in a word and are related with underlying syllables' durations.

Index Terms: Speech production and perception, tone languages

1. Introduction

Syllabic tones in tone languages are connected with distinct *F0* patterns (rising, falling etc.). Mandarin has four different lexical tones: high (1), rising (2), low (3), and falling (4) (commonly used tone indices are given in brackets). A so-called neutral tone (0) lacking a clear tonal target occurs in certain function words. Research has shown that these tone contours patterns can be associated with distinct *F0* patterns which are strongly influenced by the tonal context (tonal co-articulation between subsequent syllables) [1], focus [2][3], as well as sentence intonation [4]. Mandarin which is written with characters corresponding to phonetic syllables has often been characterized as a monosyllabic language, but in reality a large number of words actually comprise two and more syllables. Prosodically speaking, a group of syllables corresponding to a compound word can contain up to five and more syllables which are uttered as a unit. In terms of phonetics, Mandarin is a relatively impoverished language with only about 400 different syllables; therefore the four tones are indispensable for creating a sufficiently large repertory of about 1300 combinations. In modern Mandarin we witness a very productive process of forming new compound words with four and more syllables. It has been shown in previous studies that citation forms of tones are more easily identified than those produced in running speech [5]. This finding suggests that longer words do not depend as heavily on the precise tonal realizations for disambiguation as mono-syllables and also raises the question whether the number of syllables in the prosodic group influences the degree of co-articulation. Since a very long word might have a smaller potential for confusions than a short one this could imply a less careful realization of tones. The current study therefore addresses this issue by extracting monosyllabic tokens from a large database in which these syllables form part of prosodic words (henceforth short 'words') containing between one and five syllables. All tokens were produced fluently and not as citation forms. Based on a statistical analysis of the words in the database we identified 10 sets of syllables which occurred with all four tones in words of varying length. In order to quantify the tonal reductions - if they occurred at all - the monosyllabic tokens were presented to native speakers of Mandarin in a perception test.

2. Speech Material

The database employed in this study contains a total of 52739 syllables and an equivalent of about 3 hours and 18 minutes of speech produced by a female speaker. It was developed for a speech synthesizer based on non-uniform unit concatenation [6] and contains 2894 sentences of news-style readings. The database was annotated on the syllable/tone level and also contains prosodic group boundaries. The latter sometimes correspond to lexical word boundaries, but more often were labeled rather on prosodic grounds, that is, taking into account the prosodic coherence/distance of subsequent syllables. In the scope of our study we are mostly interested in

the effect of the super-ordinate prosodic group (which can contain more than one lexical word) on the individual syllable. For example, the word 'ren min ri bao' (*People's Daily*) actually contains four lexical words (ren - *man*, min - *the people,* ri - *day/sun*, bao - *journal*), more reasonably we would interpret it as consisting of two lexical words (renmin - the people, ribao - newspaper), but it is uttered as a single prosodic unit, therefore we regard it as a four-syllable word. Statistical analysis of the database yielded the results presented in Table 1.

As could be expected the predominant pattern is disyllabic, accounting for almost 64% of prosodic words whereas tri-syllabic words come out second. Most of the five-syllable units are actually four-syllable groups followed by 'de' like "ren min ri bao de"-*"of the People's Daily"*. Judging from the mean syllable durations we can say that syllables belonging to larger units are more strongly compressed.

Table 1: *Statistics of database. Frequency and percentage of prosodic groups with 1-5 syllables and mean syllabic duration.*

Number of syllables	frequency	Percentage	Mean syll. duration [ms]
1	2508	10.4	258
2	15345	63.7	228
3	5471	22.7	218
4	702	2.9	206
5	64	0.3	196
total	24090	100	225

In order to examine whether the size of the super-ordinate group has also an influence on the tonal realization we designed a perception experiment with material taken from the database. We assumed that if more tonal reductions occurred in larger words they should lead to a reduced rate of tone recognition once a syllable was excised from its context and presented in isolation. It should be stressed that even the monosyllabic words in the database were spoken fluently and did not correspond to citation forms.

We subsequently searched the database to find sets of syllables that met the following requirements:

1. they occurred as meaningful monosyllabic words
2. they occurred with all four tones
3. they occurred in a large number (> 200) of polysyllabic words of varying length and in varying positions

Finally we chose the following ten syllable sets, denoted in Hanyu Pinyin:

da	gong	ji	jia	shi
xian	xiang	xing	yi	yu

We did not include syllables with neutral tone 0 because their tonal contours largely depend on the surrounding context and are therefore difficult to reliably identify on an isolated syllable.

In order to choose appropriate tokens for the perception test covering a range of conditions we examined the total of 3595 words containing one or more of the ten syllables and selected 564 words by the following criteria: If a condition, say, the syllable *da* occurring with tone 2 as the first syllable in a disyllabic word, was represented by only one word we selected that word, if there was more than one example we sorted the words by the respective durations of the target syllable and selected the one at the median position and the one next to it, hoping to avoid extremely long or short exemplars. This strategy was used to create three lists of stimuli, each containing 214 items, that is, a total of 642. This means that some of the 564 words were used more than once.

A trained phonetician examined the prosodic words chosen to confirm that they had been correctly annotated and were meaningful, and that the target syllables carried the tone they were supposed to. Furthermore she examined whether the target syllables were light or heavy. Although a comprehensive theory of word stress in Mandarin remains yet to be developed [7], syllables can assume different weights depending on the phrasal context. Syllables carrying the neutral tone, for instance, are invariably pronounced light. For tones 1-4 the weight is commonly believed to be mainly influenced by the syllabic structures of syllables making up a prosodic unit. We found that only 15 tokens in the polysyllabic words chosen for the perception test corresponded to light syllables.

3. The Perception Test

Experiments were conducted using the *DMDX* software [8] and employed scripts provided by Caroline Jones (MARCS, University of Western Sydney, Australia) that were slightly modified. Considering the large number of stimuli and the fact that the tests were to be conducted with native Mandarin speakers, an identification task rather than a discrimination task was employed. This required participants to identify the presented stimulus by choosing one of four syllables written in Pinyin, for instance, da1, da2, da3, or da4.

Some of the syllables were chosen for a practice session preceding the experiment proper, and the 214 stimuli in one trial were divided into two groups of 107 divided by a short break. Within each of the three different trials tokens were presented in randomized order and each trial took about 20 minutes to complete.

Participants listened to the stimuli over headphones connected to a PC soundcard. Each trial started with a preparation phase of one second during which the word 'ready' was displayed. Then the stimulus was presented, followed by the four Mandarin syllables. The order of the syllables corresponded to the numbering conventions for Mandarin tones and was left unaltered during the experiment. Following the presentation of the syllables, participants made a forced choice by hitting the appropriate number key on the keyboard.
In the practice trials, feedback concerning response accuracy was given, but in the main test not.

Participants were 15 members of staff (15 female) at iFlyTek corporation, Hefei, China, aged 23-30. They reported to have normal hearing and were trained in speech annotation tasks. Most of them were familiar with the speaker who had produced the speech data. Each of the three trials was performed by five subjects.

4. Results

An earlier study on the perception of Mandarin syllables uttered in isolation in various audio-only and audio+video conditions [9] will serve as a benchmark for comparing the results of the current experiment. Table 2 displays the percentage correct and the corresponding confusion matrix on clean audio. As can be easily see, the recognition rate is close to 100% and the largest though negligible number of confusions occurs between tones 2 and 3.

Table 2: *Proportion correct and confusion matrix on citation forms in percent taken from [9].*

Tone	1	2	3	4
percent correct	99.6	96.9	99.4	99.6

intended tone	perceived tone [%]			
	1	2	3	4
1	99.6	0.4	0	0
2	0.8	96.9	2.3	0
3	0	0.6	99.4	0
4	0.4	0	0	99.6

In contrast, the pooled result of the current perception experiment showed the picture displayed in Table 3. As can be seen, recognition rates drop considerably, especially for tone 3 which is often confused with tone 4.

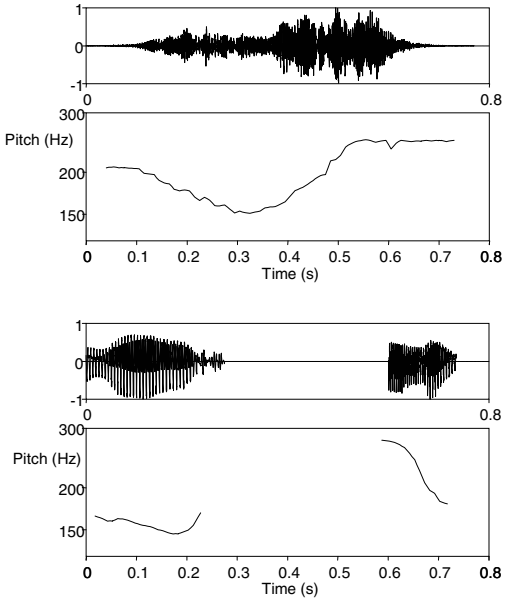

Figure 1*: Examples of syllable yi carrying tone 3, waveforms and f0 contours. Top: Citation form, bottom left: phrase-final, bottom-right: word-medial.*

In terms of tonal configuration, citation forms of tone 3 exhibit a falling-rising *F0* pattern (see example in Figure 1, top, analyzed and plotted using *PRAAT* [10]). When inserted into the context of a phrase tone 3 may partly lose the rise when it occurs phrase-finally (Figure 1, bottom left), or the rise is delayed into the following syllable, making it harder to identify (Figure 1, bottom right). The figure nicely illustrates the dramatic changes tone 3 undergoes in a phrasal context. Still, the example at the bottom left was consistently identified, presumably due to the vocal fry at the end of the syllable which is a well-known feature of tone 3.

Table 3: *Proportion correct and confusion matrix, pooled result of current perception experiment.*

tone	1	2	3	4
percent correct	93.5	79.1	49.0	82.2

intended tone	perceived tone [%]			
	1	2	3	4
1	93.5	2.0	1.3	3.2
2	8.5	79.1	10.1	2.3
3	15.7	5.2	49.0	30.1
4	8.8	1.8	7.2	82.2

The 15 syllabic tokens classified as 'light' reached a chance-like mean recognition rate of 17.5% and had apparently lost their tonal specification altogether.

Table 4: *Proportion correct depending on the size of the super-ordinate word and number of stimuli N for each type of word (Σ=564).*

syllables in word	1	2	3	4	5
percent correct	80.2	76.5	76.2	75.6	66.8
N stimuli	53	150	210	131	20

When we examine the relationship between correct responses and the length of the prosodic word, we find that mono-syllables are identified more reliably than poly-syllables, and this result is significant ($p<.05$) but the number of syllables does not seem to clearly influence the correctness of responses (see Table 4), although the result for five-syllable words from which we had very few tokens suggests that the recognition rate might further decrease as the prosodic group expands.

We subsequently checked whether the position in a prosodic group influenced the recognition rate. Since there were relatively few tokens from five-syllable words, Table 5 only displays the results for two- to four-syllable words. The figures suggest that syllables in group initial position are more reliably identified than in the post-initial position. In three- and four-syllable words the final syllable yields a relatively high recognition rate. As can be seen from the number N of stimuli not all syllable positions were equally frequent in the stimulus material. We suspected that the high recognition rate four-syllable-word-finally was due to the fact that such a large unit often occurs at the end of a prosodic phrase which is subject to final lengthening. Since tonal pattern and syllable duration are closely linked this relationship might also explain the other results. We therefore examined the mean syllable durations of the whole database depending on the syllable position (see

Table 6). The figures suggest a certain correspondence for three- and four-syllable words, but the situation is reversed in the disyllabic case. The final-lengthening effect, however, is obvious.

Table 5: *Proportion correct depending on the position of the syllable in the super-ordinate word.*

syllables in word	position	Percent correct	N stimuli
2	1	80.2	79
	2	72.3	71
3	1	81.2	74
	2	71.6	70
	3	75.6	66
4	1	77.0	36
	2	70.0	40
	3	72.0	34
	4	89.4	21

Table 6: *Mean syllabic duration depending on the position of the syllable, calculated for the entire database.*

Syllables in word	position	Mean duration in ms
2	1	215
	2	241
3	1	208
	2	190
	3	255
4	1	205
	2	181
	3	195
	4	241

We calculated correlations between the proportion correct and the syllabic duration of the tokens used in the perception experiment and found ρ=.20 (p<.01). If we exclude tokens that were never misclassified this figure rises to .24. This outcome suggests that the main factor influencing the recognition rate besides the tone of the syllable is the syllabic duration. The length of the super-ordinate group (except for the mono-syllable/poly-syllable distinction) does not have a tangible influence, at least in the data examined. The results from the perception test and the duration analysis, however, suggest certain rhythmic patterns in poly-syllables which might indeed affect the tonal realization. As could be expected the reaction time of the subjects is negatively correlated with the proportion correct (ρ=-.24, $p < .01$) and amounts to an average 1.48 s for tokens from monosyllabic words and 1.71 s for those from poly-syllabic words.

5. Discussion and Conclusions

The current study investigated the relationship between the size of a prosodic group and the tonal realization. Syllabic tokens were excised from their original context and presented in a perception test. The outcome of the test showed considerably lower recognition rates than for citation forms, especially for tone 3 which was frequently confused with tone 4. Monosyllables on the average were 5% more often correctly classified than syllables excised from polysyllables, but the number of syllables in the polysyllabic word did not seem to significantly influence this rate. However, the reduction of the recognition rate appears to depend on the position in the prosodic group which obviously also strongly influences the duration of the syllables.

It must be noted that the database we used had not been specifically created for the purpose. As a consequence, we did not yield a perfect balance in the stimulus material regarding tone, size of word, and position in word. Furthermore, the reading style material that had been recorded for a speech synthesis application will certainly not exhibit as strong tonal reductions as could be found in spontaneous speech.

Correctly labeling prosodic groups in Mandarin proves to be problematic, as on the one hand it is guided by the lexical decisions of the labeler and on the other hand by the actual performance of the speaker. This situation is aggravated by the fact that Mandarin is written without spaces between words. For nine stimuli initially classified as pertaining to monosyllabic words, for instance, the prosodic grouping had to be revised as they belonged to larger units. We cannot exclude that the result of our experiment was flawed by these circumstances. In the future we plan to perform a similar experiment on better controlled and balanced data, as well as examine more spontaneous speaking styles that might be more prone to tonal reductions.

6. References

[1] Xu, Y. "Consistency of tone-syllable alignment across different syllable structures and speaking rates." *Phonetica* 55: pp. 179-203, 1998.

[2] Xu, Y. "Effects of tone and focus on the formation and alignment of F0 contours." *Journal of Phonetics* 27: pp. 55-105, 1999.

[3] Chen, G., Hu, Y., Wang, R. and Mixdorff, H. "Quantitative Analysis and Synthesis of Focus in Mandarin", in *Proceedings of TAL 2004*, pp. 25-28, Beijing, 2004.

[4] Shih, C. "Tonal effects on intonation", in *Proceedings of TAL 2004*, pp. 163-167, Beijing, 2004.

[5] Xu, Y. "Production and perception of coarticulated tones", *JASA*, 95: 2240-2253, 1994.

[6] Ling, Z., Hu, Y., Shuang, Z., Wang, R. "Compression of Speech Database by Feature Separation and Pattern Clustering Using STRAIGHT," In *Proceedings of ICSLP2004*, pp. 1201-1204, Cheju, Korea, 2004.

[7] Deng, D., Chen, M, and Lu S. "Study on Stress Models of Chinese Disyllable", in *Proceedings of TAL 2004*, pp. 49-52, Beijing, 2004.

[8] http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm

[9] Mixdorff, H., Hu, Y. and Burnham, D. "Visual Cues in Mandarin Tone Perception." In *Proceedings of Eurospeech 2005*, pp. 405 - 408, Lisbon, Portugal, 2005.

[10] http://www.praat.org